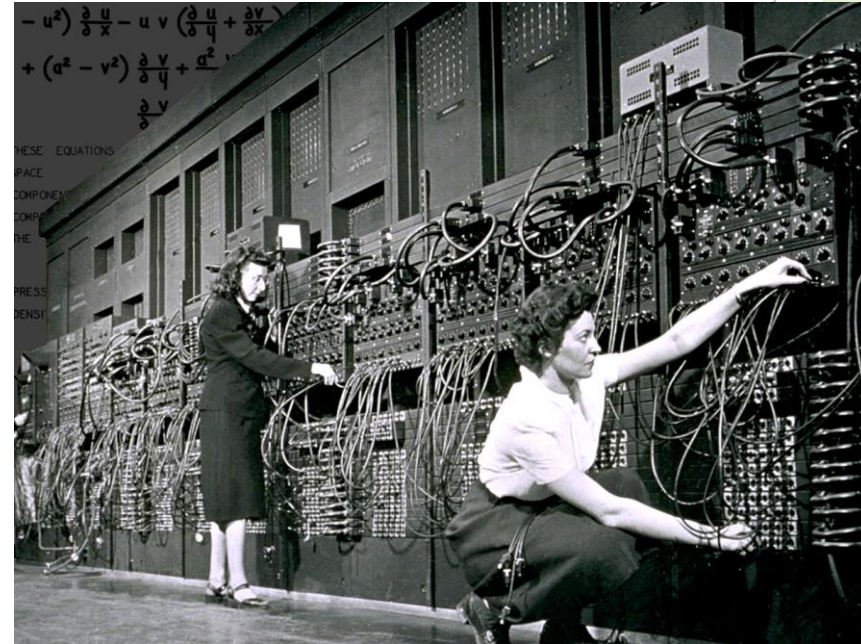
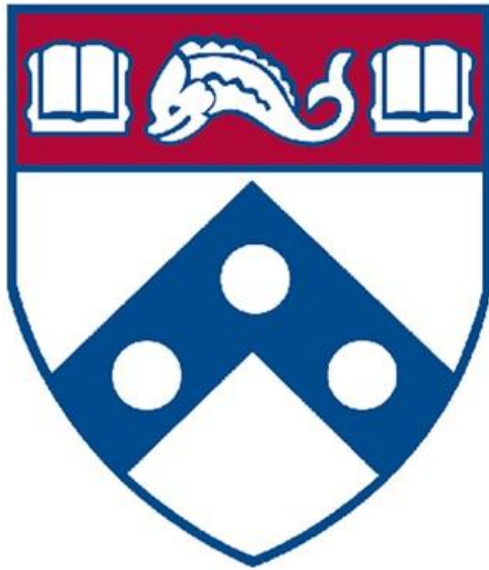


Text Generation and Sentiment Analysis of College Confidential

What are the Chances?

- **Objective:** to build a text generator on a corpus of forum posts





The Data

- ▶ 16451 comments from the University of Pennsylvania's College Confidential Forum
 - ▶ Somewhat formulaic
 - ▶ VERY opinionated
- ▶ Scraped using Beautiful Soup
- ▶ Preprocessed to get rid of capitalization, punctuation, usernames, etc.



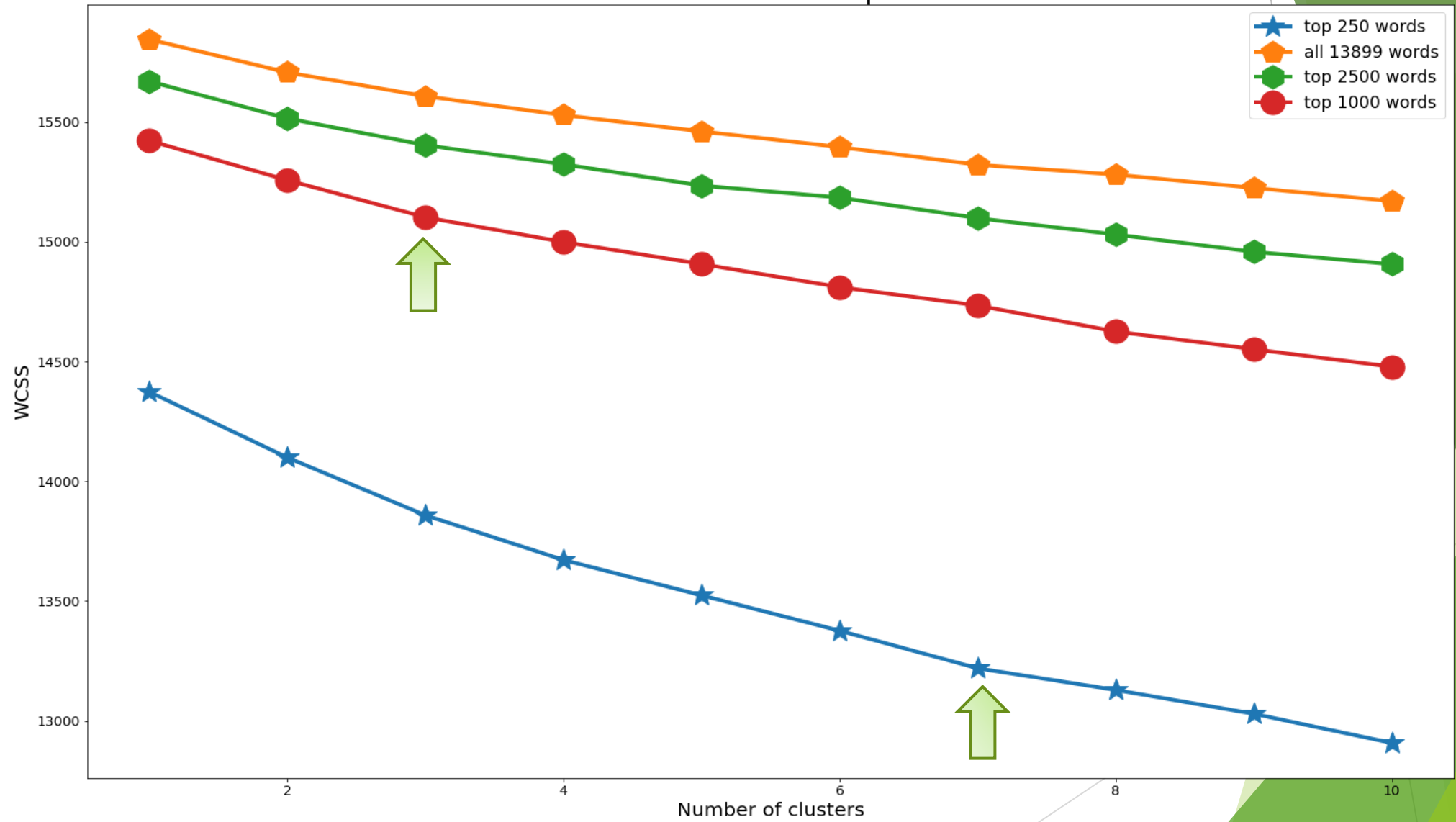
I think you are below average stat wise for almost all of the schools. The ones mentioned by tR5674 or at best average.

Your ECs are average to slightly above for most of these schools.

Your demographic doesn't help you either. Being an upper class white kid gets you no bump.

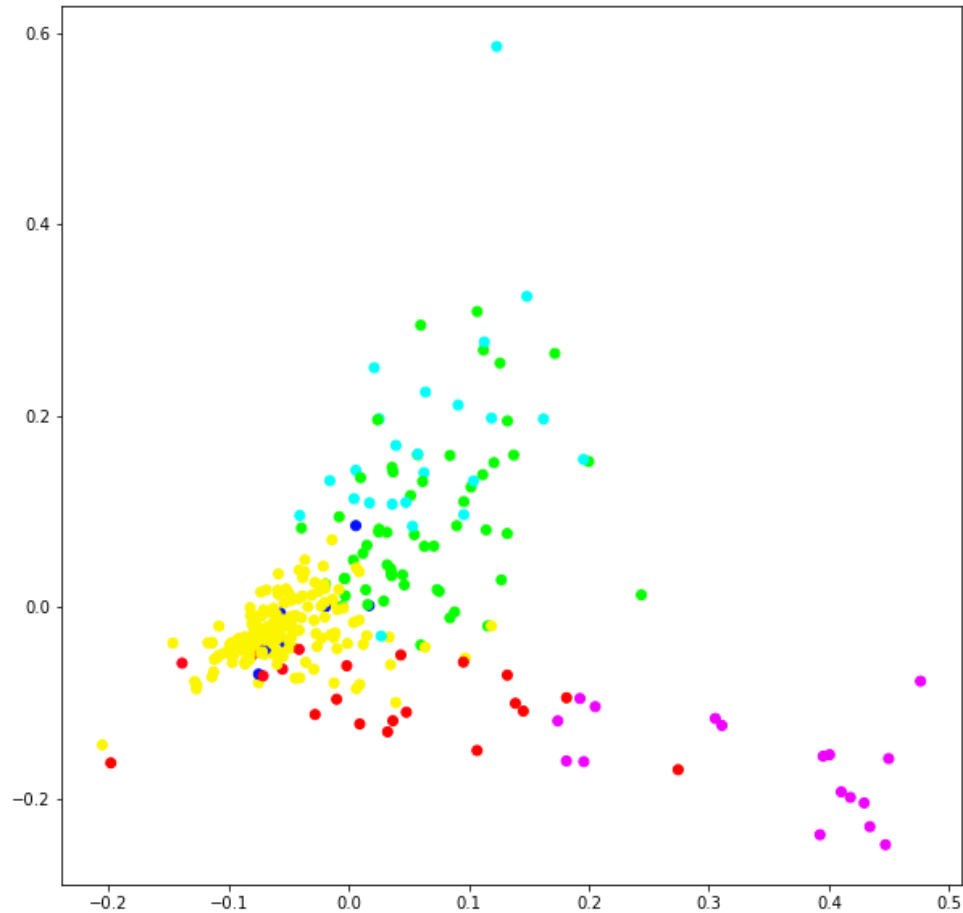
I do know you won't get accepted unless you apply, but your chances are low.

The Elbow Graph

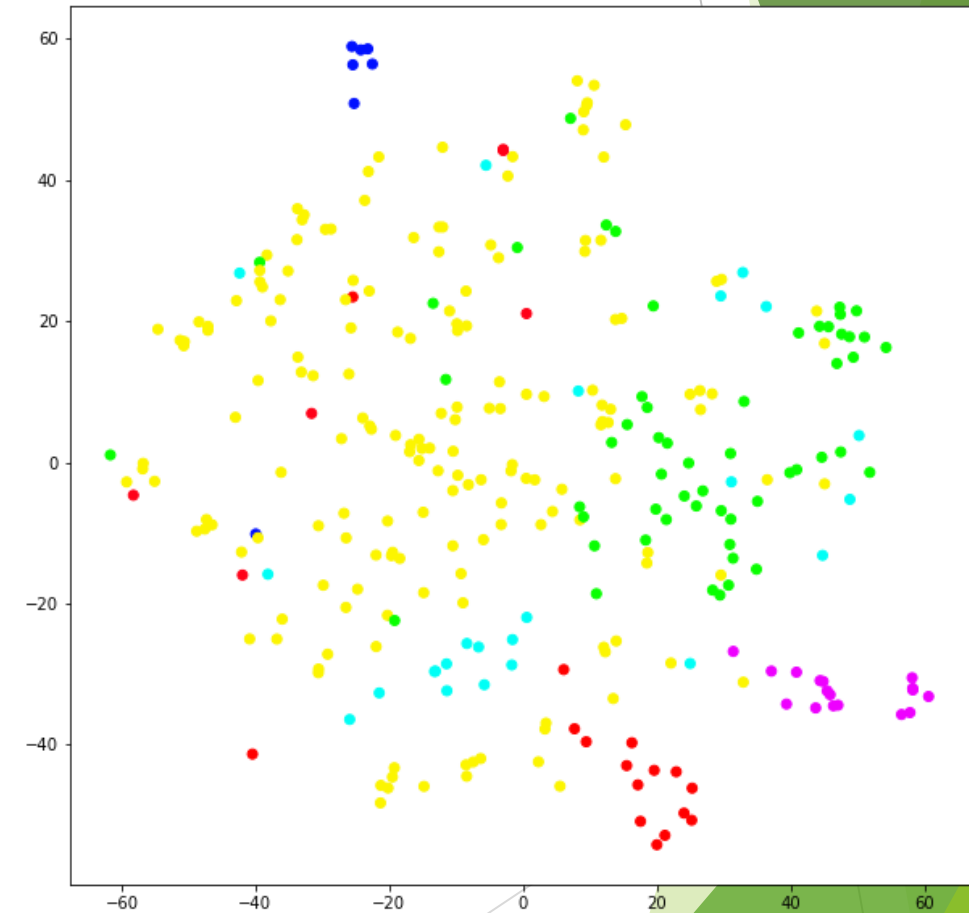


Visualizing the Data

PCA Cluster Plot



TSNE Cluster Plot



Is the Data separable?

- ▶ Clustering suggests separability
 - ▶ Couldn't easily discern a “theme”/format
- ▶ Topics may overlap too much because of the forum's narrow focus

So, let's try something a bit more complex!



Keras (only a few epochs)

- ▶ Human: “ I am applying to Penn ED. What are my chances? My SAT score is 1500.”
- ▶ Computer: “You are a strong applicant with a high SAT score”

Fails Turing test!

Improvements

- ▶ More time!
- ▶ Model that can reply in real time
- ▶ More cleaning and tagging of ideas/synonyms
- ▶ Revisit less intensive models for semantic analysis

Thanks!

Any Questions?